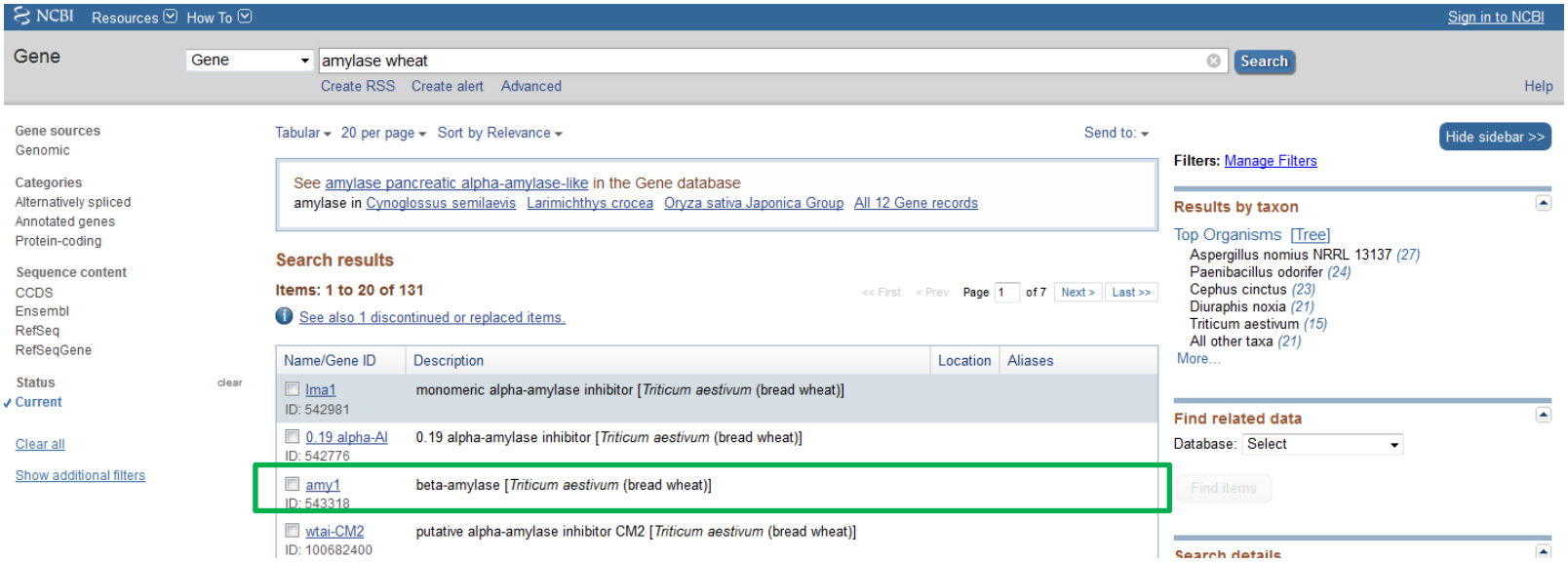
Practice I

Identify a candidate gene from another genome and extract the sequence (including flanks)

Find and upload candidate gene sequence and reference

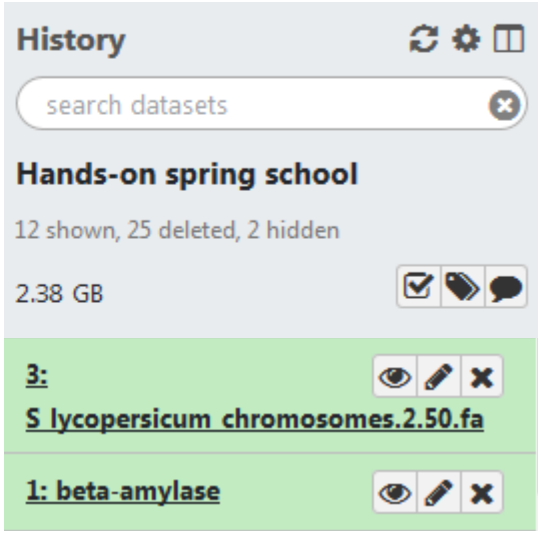
- Look for the term ‘Amylase wheat’ on https://www.ncbi.nlm.nih.gov/



- Export in fasta format

Find and upload candidate gene sequence and reference

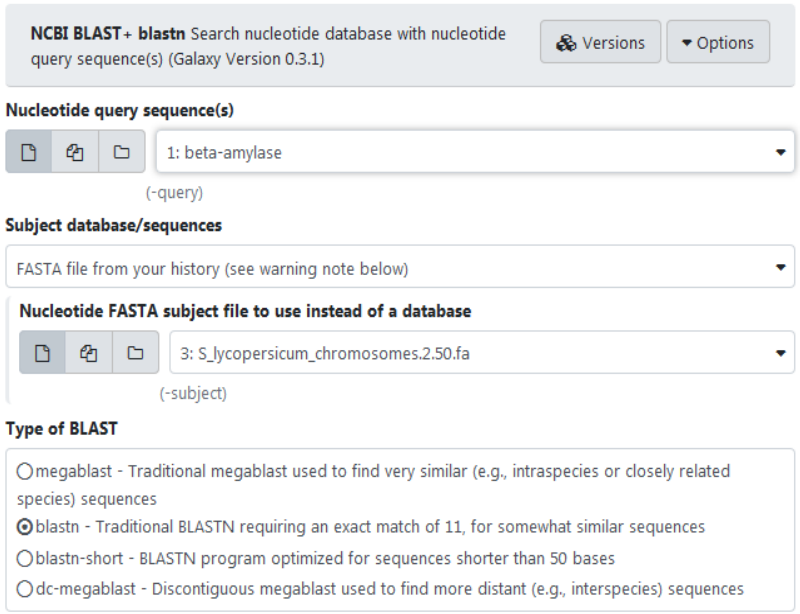
- Get tomato reference from the ‘Galaxy workshop’ folder
- Your history should contain the following files now:



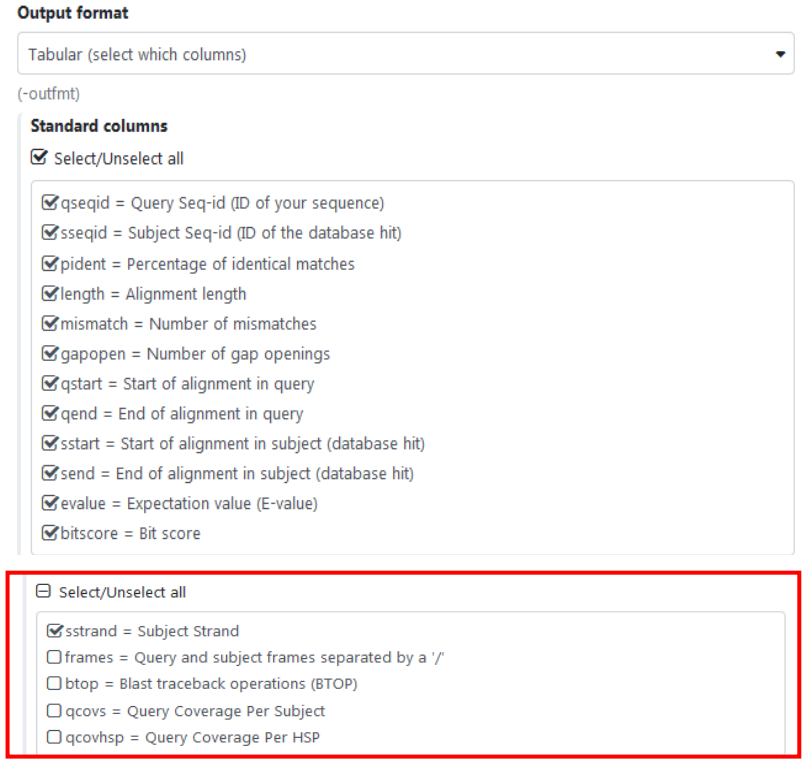
BLAST your sequence

- Identify the location of the beta-amylase gene in the tomato genome

Input



Don‘t forget to choose ‚sstrand‘ option for the output

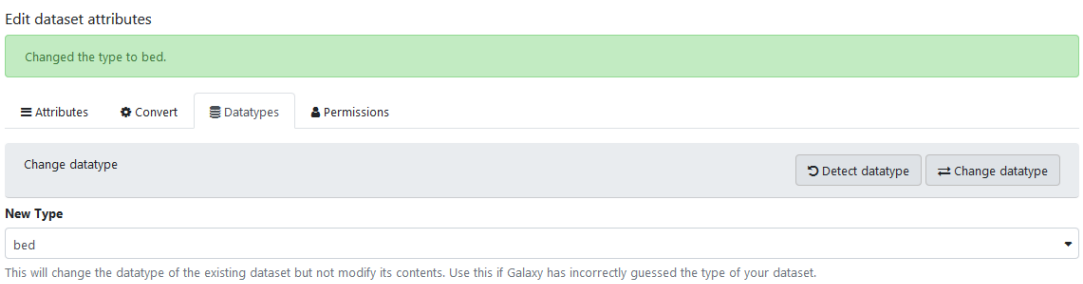


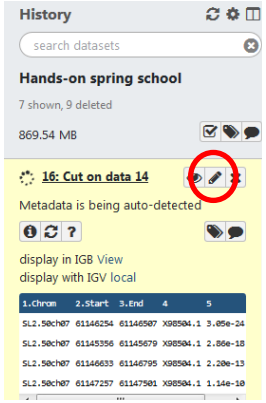
Prepare BED file from BLAST result

- Cut columns from a table ~~(cut)~~ !choose correct tool!
  - 2, 9, 10, 1, 11, 13 (sname, sstart, send, qname, score, sstrand)
- Table should look like this:

1	2	3	4	5	6
SL2.50ch07	61146254	61146507	X98504.1	3.05e-24	plus
SL2.50ch07	61145356	61145679	X98504.1	2.86e-18	plus
SL2.50ch07	61146633	61146795	X98504.1	2.20e-13	plus
SL2.50ch07	61147257	61147501	X98504.1	1.14e-10	plus

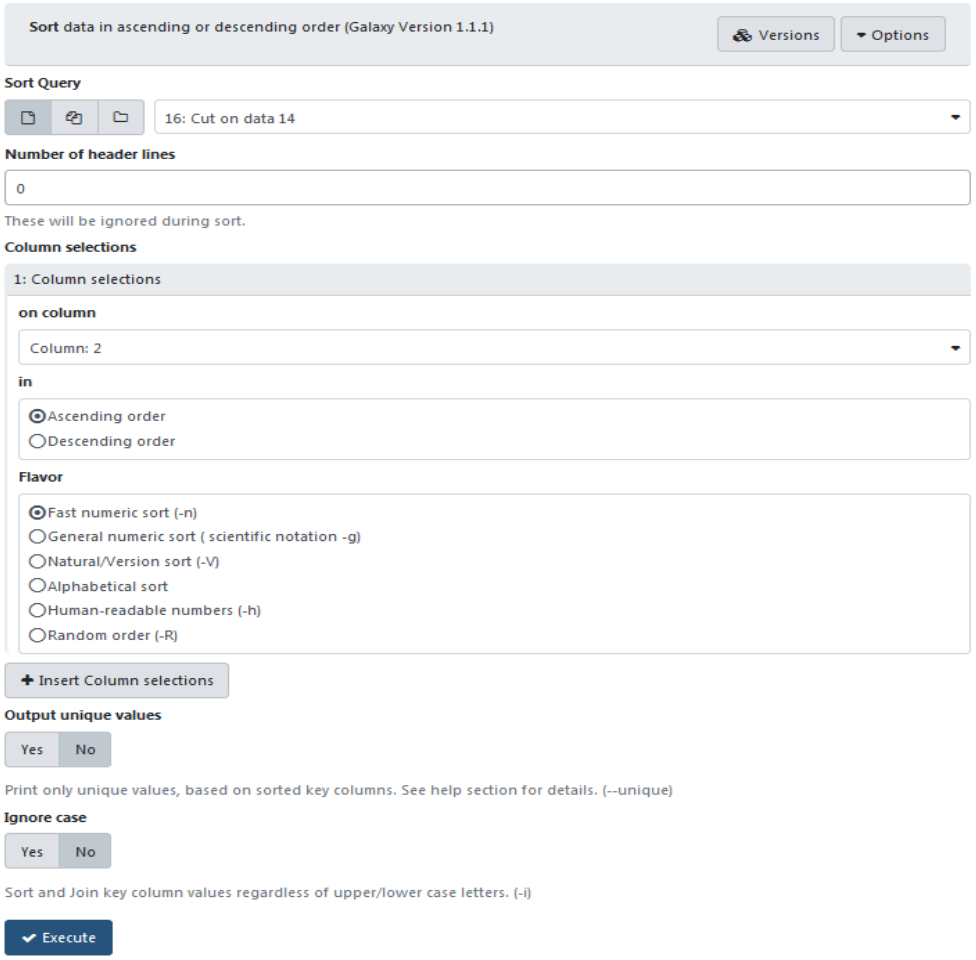
- Convert ‚tabular‘ to ‚bed‘ format





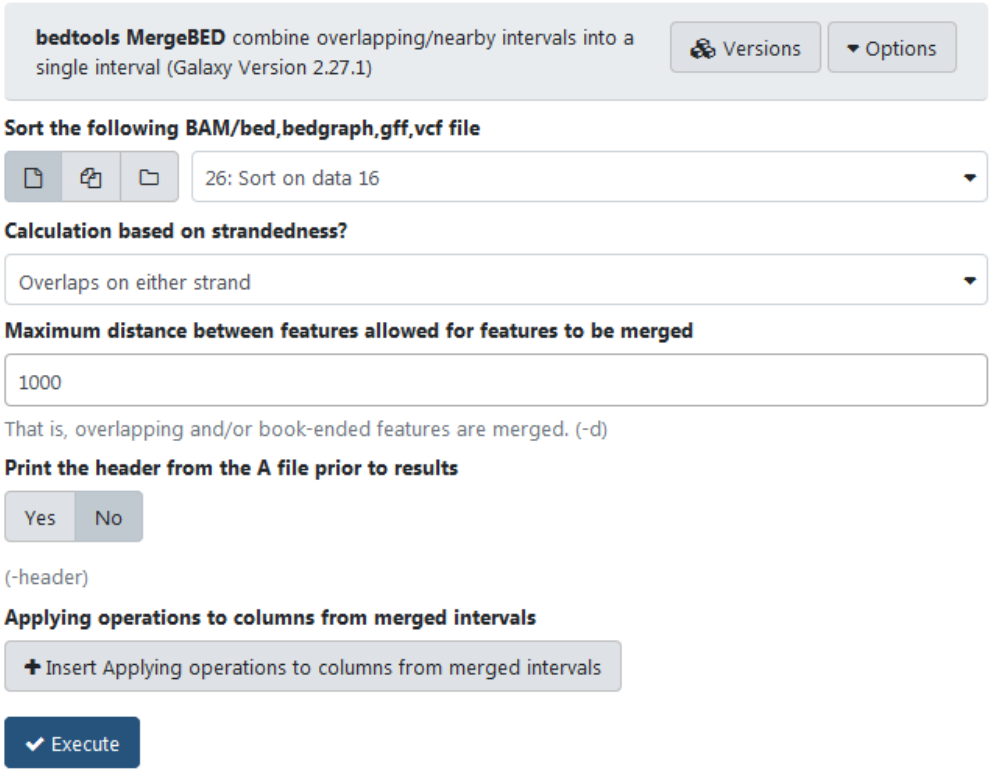
Sort the tomato coordinates

- Sort data in ascending or descending order
  - Required for bedtools merge
- Optionally, bedtools SortBED can be used



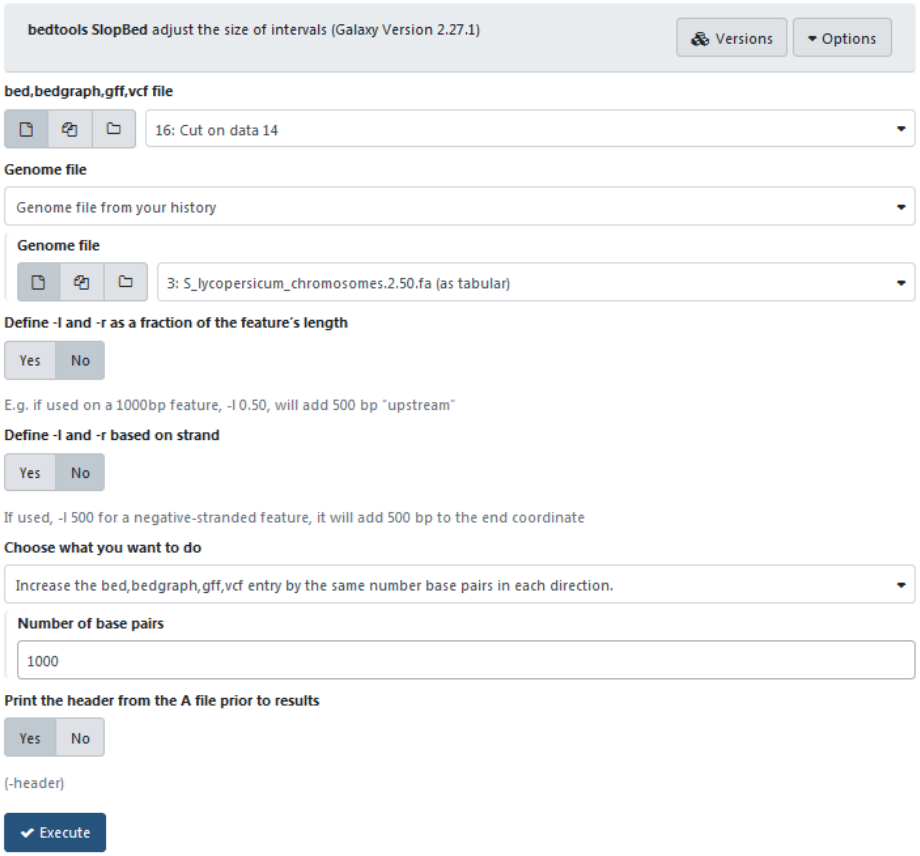
Merge the exons into a single feature

- bedtools MergeBED combine overlapping/nearby intervals into a single interval



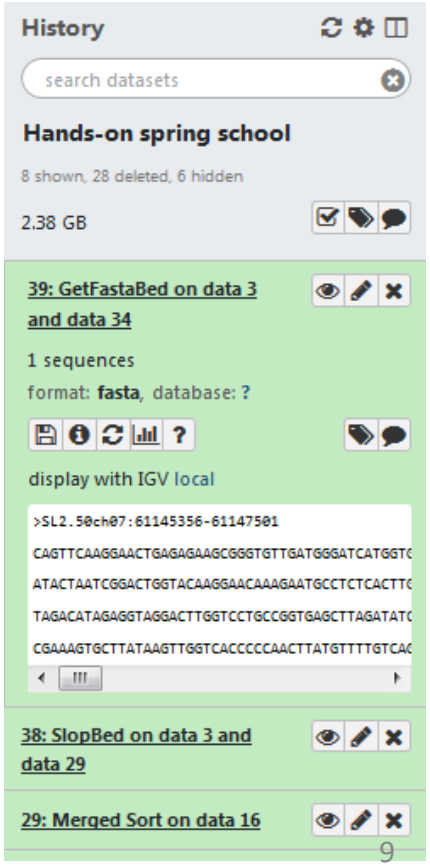
Expand flanks to cover the whole gene

- bedtools SlopBed adjust the size of intervals



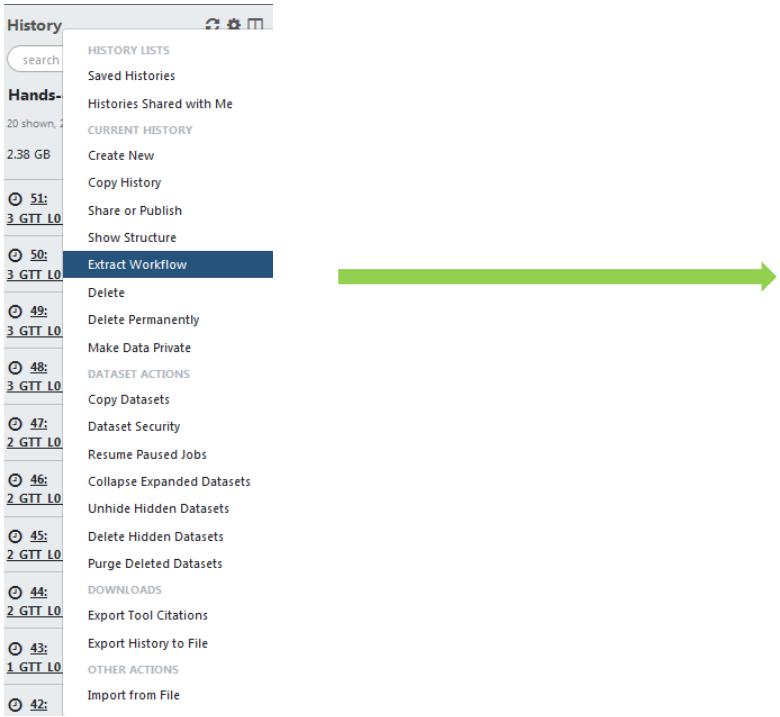
Finally, extract the sequence

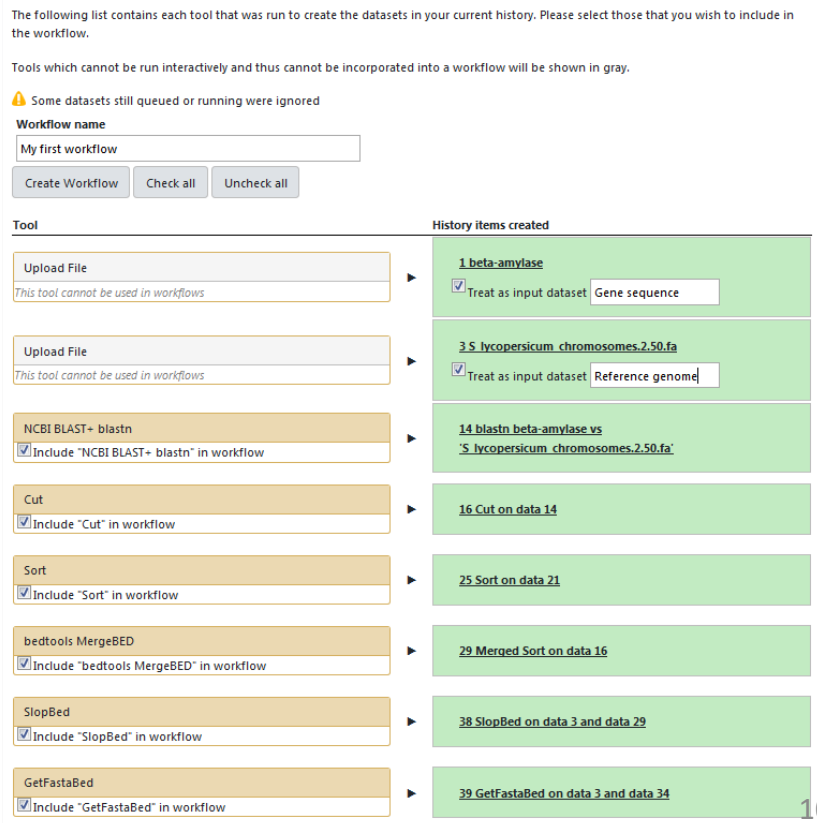
- bedtools GetFastaBed use intervals to extract sequences from a FASTA file



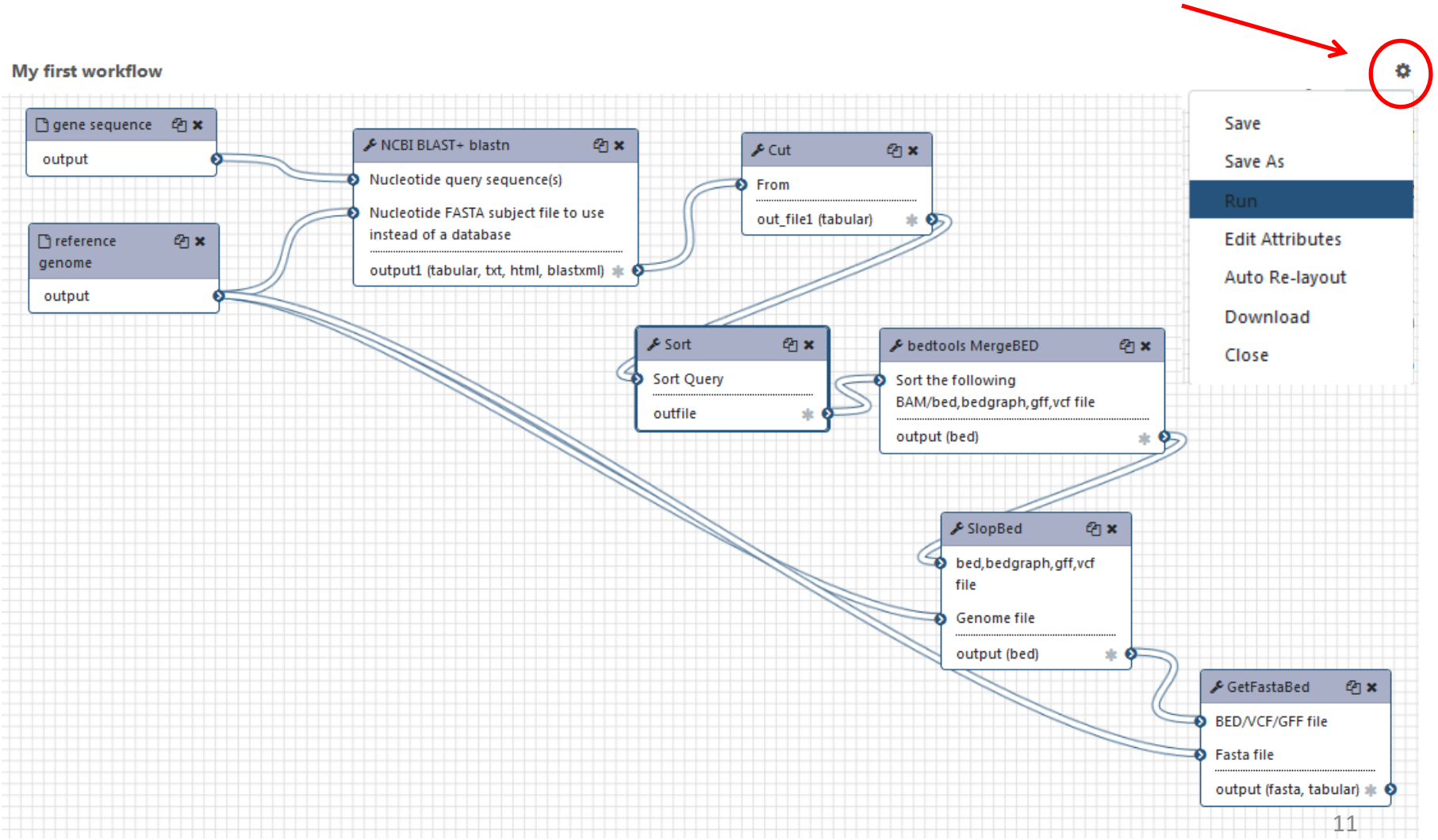
Extract a workflow

- Use the ‚Extract workflow‘ button in your history options
- Change name of the two inputs and the workflow itself – ready



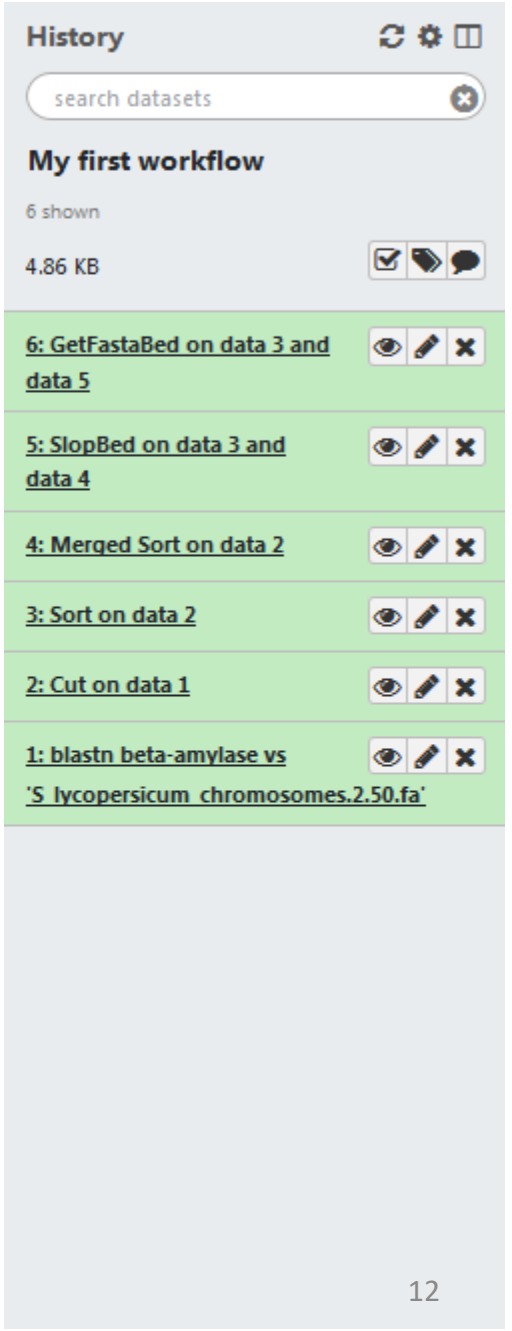


- Some correction might be required, if so, please re-order and connect to tools
- If you have fixed the tool order, you can save and test the workflow by using the gear symbol in the right upper corner



- After you have started your workflow, the resulting history should like this

- First task completed

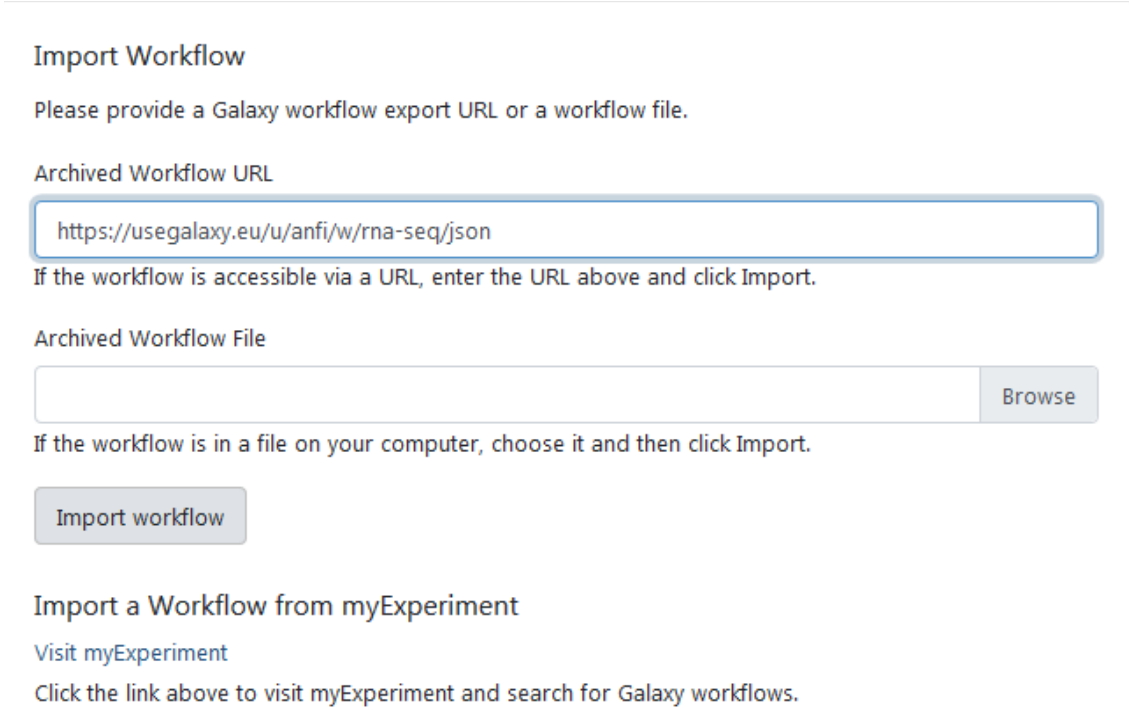


Practice II

Dataset collections and RNA-Seq workflow

Import workflow

- Add the following adress
  - https://usegalaxy.eu/u/anfi/w/rna-seq/json

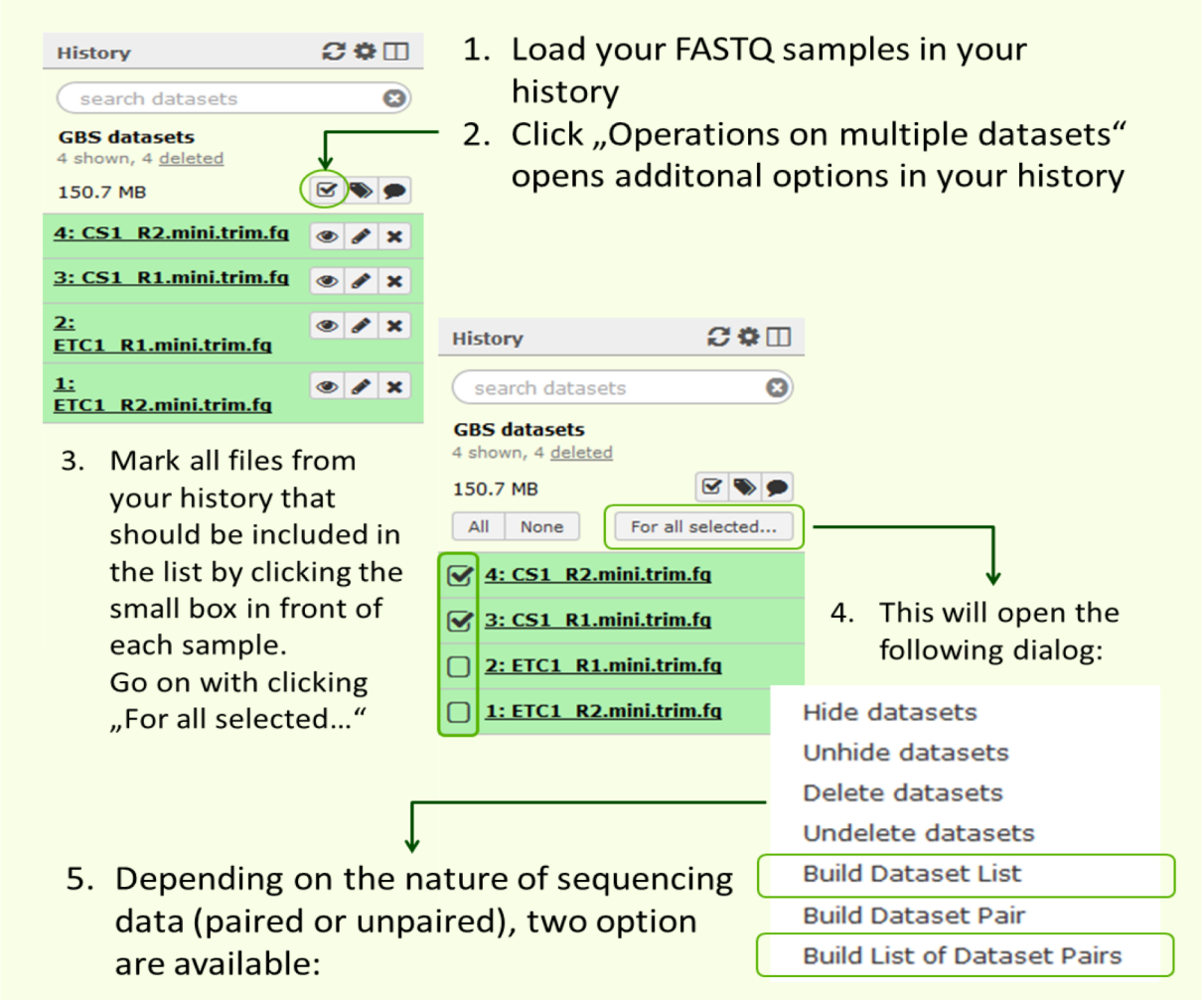


References and input data

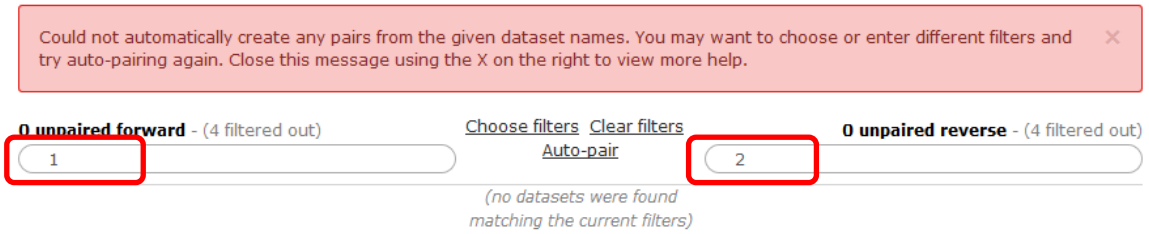
- Tomato genome sequence is already in your history (first exercise) -> please copy it to a new history
- Import tomato GFF file (link can found in couse material)

  ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG2.4_release/ITAG2.4_gene_models.gff3
- Upload all fastq files from the folder ‚rnaseq_rawdata‘

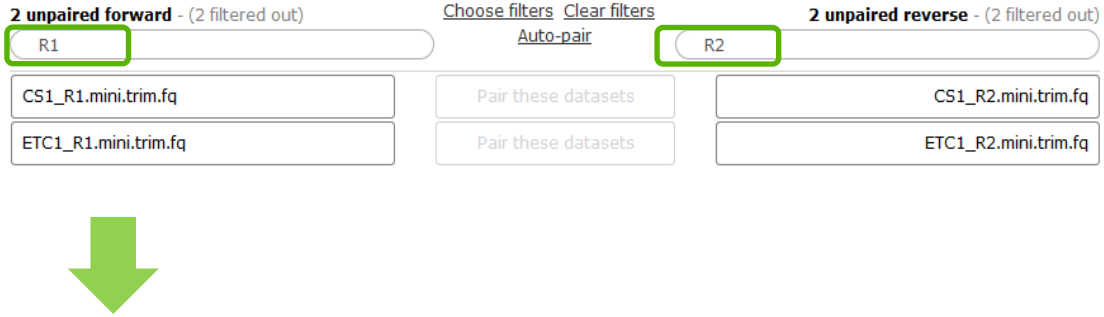
Creating dataset collections



Note: If your pairs are not named by „_1“ and „_2“ ending, you will probably see this message:

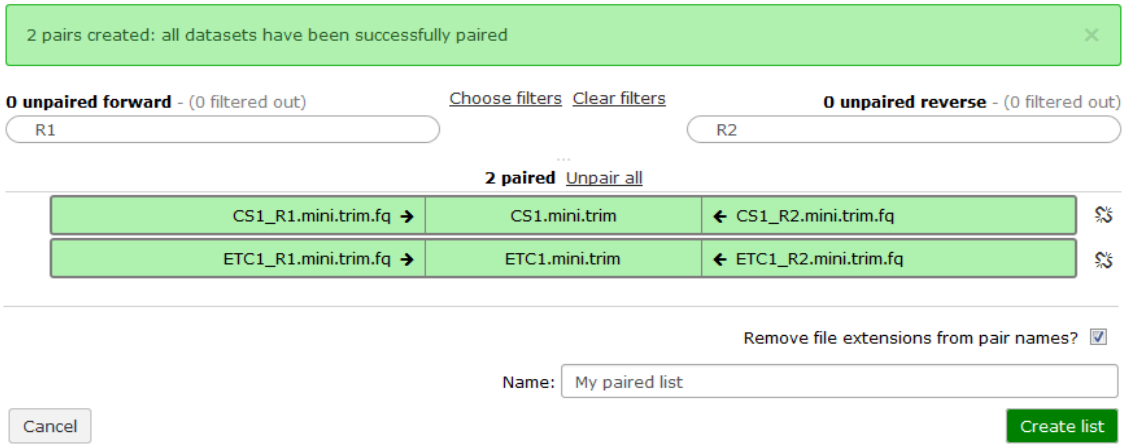


This should be corrected by typing the naming pattern of your data.
In this example „_R1“ and „_R2“:



Check the correct pairing of the files. Then click „Auto-Pair“.

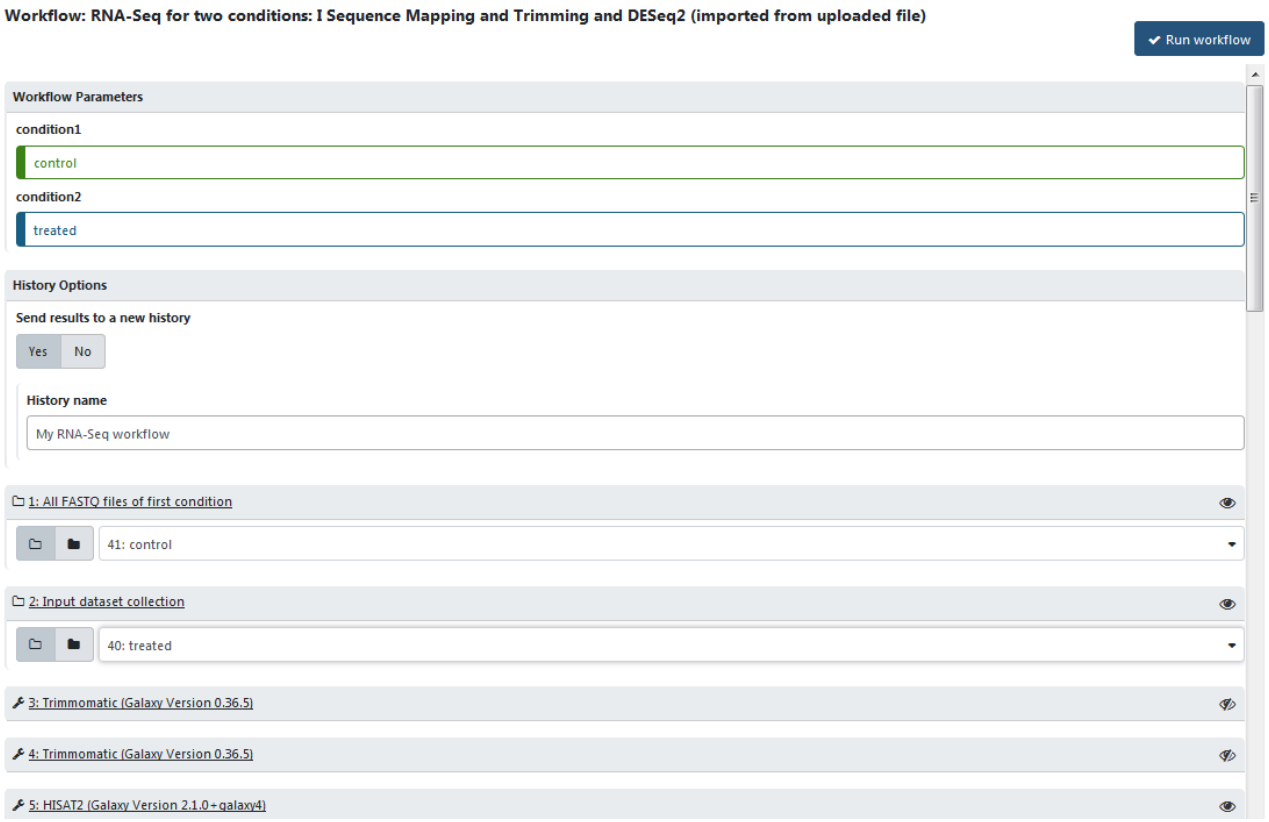
The files are paired according to their name scheme. Name and create the list now.



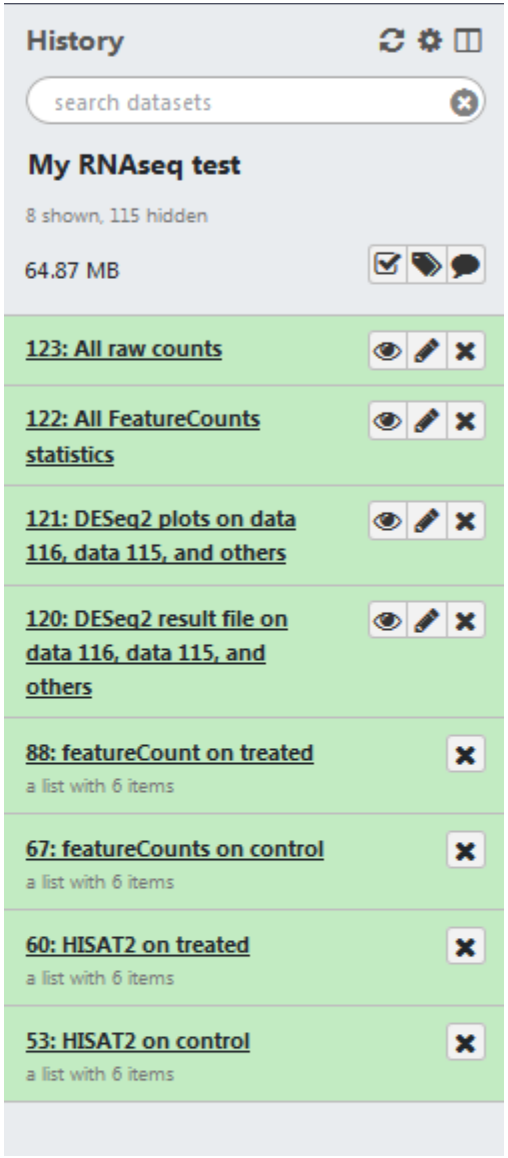
- Your history should now show two collection, each containing 6 paired-end files
- Everything is prepared to start the workflow

Starting the workflow

- Go to ‚Workflow‘ menu and select ‚Run‘ for the RNA-Seq workflow
- Check whether the ‚GFF gene identifier‘ option (Advanced option) is set to ‚Parent‘



- This is the final history (right)
- Please inspect the results
- Improvements:
  - Renaming and joining the samples previously
- Second task is completed



Bonus

- Please edit the workflow to include
  - Fastqc to check read quality
  - a clipping tool (e.g. Trimmomatic)

  before the HISAT2 mapping step